We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

3

Data Mining Techniques for Explaining Social Events

Krivec Jana and Gams Matjaž
Jožef Stefan Institute, Slovenia

1. Introduction

When trying to discover patterns and classification models for social events, machine learning can be a powerful tool. The most common usage of data mining techniques is categorization of new examples into specific classes. Nevertheless, the simple number indicating classification accuracy, as with SVM or similar non-transparent methods, is usually not good enough for the case when we want to understand the problem and check the obtained relations with human sense and knowledge. It is not good enough because we don't know whether the relations beneath are logical to human experts of the research field or even we don't know how the relations look like. We want to check the computer-constructed relations whether already known or created anew. In many cases when dealing with social events, it is of extreme importance to combine computer and human knowledge. Classification trees or classification rules seem to be the best choice for this kind of problems. The problem that might arise in this case is in the quality of the discovered patterns, e.g. it is well known that some computer-generated relations seem to be important, but statistically do not exceed the chance of random choice. That is why the procedure of conducting the best possible classification trees or rules from the data must follow certain rules. To put it shortly, first, data has to be manipulated in various, yet systematic ways in connection with opinions of a field expert. The manipulation can be executed on the level of instances, attributes, class or parameters of the data mining algorithm. Second, the quality estimation should be calculated in various ways, thus providing possibility to choose the best tree of all. We performed demographic analysis in the proposed way, obtaining some new and confirming some already published relations.

2. What to expect

In this chapter a description of the specific problem type is presented. First, a current state of the procedure dealing with social events is described, including possible inaccuracies. In the second section, a case study showing an example of the usage of machine learning techniques for mining social events is presented, discussing possible problems and future improvements.

3. Mining social events with data mining techniques

Machine learning and lately data mining are among the most successful application areas of artificial intelligence.

Whenever there are lots of learning examples, these systems learn properties of the domain and make predictions about future cases. The most common usage of DM techniques is categorization of new examples into specific classes. However, the weakness of state-of-the-art machine learning algorithms achieving the best results in terms of accuracy (e.g. ensemble methods or SVMs) is their inability to explain their predictions. There is no guarantee that these justifications will be understood by experts and other users. The induced models are often strange to the domain experts as they present the problem in a different way. Domain experts and users often perceive these methods and their predictions as black boxes.

However, machine learning is often used to get a better understanding of the relation between inputs and outputs (Mitchell, 2006). In such cases, it is usually preferable to use methods like decision trees, rule-based models, or linear models that construct knowledge which can be presented in the form of readable, understandable trees, rules and other representations thus enabling further study and fine tuning. In this way, the causes of investigated phenomena can also be discovered and described, as we might see in the case study section.

Yet, even human-transparent models often do not provide true understanding of the constructed model and can even provide counter-intuitive solutions. The induced correlations sometimes seem illogical or simply strange to the domain experts as they would explain the same case using different terms. Pazzani (1991) showed experimentally that people will grasp a new concept more easily if the concept is consistent with their knowledge. More elaborate studies of understanding new concepts were produced by the cognitive learning community (Angehrn & Gibbert, 2008). They showed that when we learn new data, we always start with our prior knowledge and try to modify it. If the new concepts are inconsistent with our prior knowledge, the new knowledge will likely be distorted or even rejected.

Data mining method are based on human-computer interaction, where user interacts with a decision tree learner to improve the trees in performance and meaning to him/her. Knowledge acquisition should be processed in a systematic way, where humans lead the data mining by comparing multiple trees constructed on different subsets of the data set and through several forms of attribute selection. By selecting the trees that are not only consistent with the data (e.g. measured by accuracy), but also meaningful, the result of the data mining process are meaningful domain models/decision trees. Such trees, in turn, contain the relations and attributes that best describe the domain.

At the end however, quantitative measures of accuracy and quality of the conducted tree should be accessed.

3.1 Conducted tree accuracy and quality estimation

In the existing literature, many different measures for evaluating the performance of information retrieval systems have been proposed. Mostly they refer to accuracy and quality of the constructed knowledge. Detailed description is presented in the next sections.

3.1.1 Classification tree accuracy estimation

Information retrieval model should contain only meaningful relations. Most meaningful relations are often considered as those with best classification accuracy. *Accuracy* is also used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition (Ivanov, 1972).

	Condition (e.g., disease) As determined by *Gold* standard			
	True		**False**	
Test outcome	**Positive**	**True positive**	**False positive**	→ Positive predictive value
	Negative	**False negative**	**True negative**	→ Negative predictive value
	↓ Sensitivity		↓ Specificity	Accuracy

Table 1. Accuracy as a measurement of positive and negative prediction values.

Accuracy can be calculated on the following way:

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{numbers of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

To estimate the accuracy of the trees, 10-fold cross-validation is often used. In this case, these are10 iterations each time taking a different single fold for testing and the rest 9 folds for training, averaging the error of the 10 iterations. The estimated accuracy of a classification tree corresponds to a probability that a new example will be correctly classified. As the best tree in particular experimental subgroup we choose the tree with the best classification accuracy (Kohavi, 1995).

Two other interesting measures are: F-measure and Under ROC Area estimation.

The F-measure is often used in the field of information retrieval for measuring search, document classification, and query classification performance (Beitzel, 2006).The traditional F-measure or balanced F-score (**F_1 score**) is the harmonic mean of precision and recall. The precision is defined as the fraction of retrieved documents that are relevant. Recall is defined as the fraction of relevant documents that are retrieved. The F-Measure is a combined measure for precision and recall (Khandefer ¸Shapiro, 2009): 2*Precision*Recall/(Precision+Recall).

Receiver operating characteristic (ROC) analysis provides tools to select possibly optimal models and to discard suboptimal ones. The machine learning community most often uses the ROC AUC (area under ROC) statistic for model comparison (Hanley, 2008). This measure can be interpreted as the probability that when we randomly pick one positive and one negative example, the classifier will assign a higher score to the positive example than to the negative. In engineering, the area between the ROC curve and the no-discrimination line is often preferred, because of its useful mathematical properties as a non-parametric statistic (Obuchowski, 2003). This area is often simply known as the discrimination. The Mann Whitney statistic is used to calculate the AUC. There are few observations regarding AUC (Fawcet, 2006):

1. It has value between [0,1]
2. A random classifier has an AUC ~0.5
3. The higher the value of AUC the better the distinguishing capability of the classifier

3.1.2 Classification tree quality evaluation

Till now, we were dealing with performance prediction. Now, we compare algorithms to see which one did better. It is not reasonable to directly use Error rate to predict which

algorithm is better as the error rate might have been calculated on different data sets. So to compare algorithms, statistical tests are needed. As indicator of tree significance we have chosen Kappa statistic, which measures the agreement of predictions with the actual class. In general, Kappa statistics are appropriate for testing whether agreement exceeds chance levels, i.e. that predictions and actual classes are correlated. The Kappa statistic measures the agreement of prediction with the true class -- 1.0 signifies complete agreement. It will usually find that predictions and actual classes are correlated and even a weak classifier will tend to show a correlation between the two (Melvile et.all, 2005). As a rule of thumb values of Kappa from 0.40 to 0.59 are considered moderate, 0.60 to 0.79 substantial, and 0.80 outstanding (Landis & Koch, 1977). Most statisticians prefer Kappa values to be at least 0.6 and most often higher than 0.7 before claiming a good level of agreement. There are also quotes that a high level of agreement occurs when Kappa values are above 0.5 and that agreement is poor when Kappa values are less than 0.3. While accuracy and AUC are correlated about 0.86, Kappa and AUC are correlated about 0.93, Kappa being the metric that is most correlated to AUC (more than MSE, Logloss/Entropy, F-measure, rank-rate or others). Kappa and accuracy, although they can give different scores, almost always make the same choices (correlation is around 0.9).

4. Case study: investigation of factors influencing country's fertility rate

For the case study we have chosen an actual problem of fertility rate being too low in some and too high in other countries. We tried to discover, which factors distinguish best the two groups of countries with different Total fertility rate (TFR). TFR is the average number of children that would be born per woman if all women lived to the end of their childbearing years and bore children according to a given set of age-specific fertility rates (Christenson et. al., 2002). If the average woman has approximately 2 children in her lifetime, this is just enough to maintain the population. Sustained low fertility rates can lead to a rapidly aging population and, in the long run, may place a burden on the economy and the social security system because the pool of younger workers responsible for supporting the dependent elderly population is getting smaller. On the other hand, too high fertility rate might cause overpopulation and other social problems. Tracking trends of fertility rates and factors that are connected with them (perhaps even influence them) helps to support effective social planning and the allocation of basic resources across generations (Gams & Krivec, 2007).
The main question of the case study is: what are the reasons for low fertility of some and high fertility of other world countries? We used data mining techniques to discover which factors differ in the countries with different TFR. Even though the idea is everything but new, the proposed approach to this problem is. So far most of the research was based on the statistical analysis. The problem is that these techniques hardly allow taking a holistic perspective (Billari et. al., 2000). Data mining techniques can overpass this deficiency.

4.1 Experimental design and procedure

We used data mining techniques as a research tool, where data were manipulated in a systematical way and results were compared with different accuracy estimation methods. Moreover, the quality of induced trees was obtained, and only the most qualitative trees were taken into further consideration.

4.1.1 Basic data description and manipulation

For machine learning and data mining, data is most commonly presented in attribute-class form, i.e. in a "learning matrix", where rows represent examples and columns attributes (Vidulin & Gams, 2006). In our case, an example corresponds to one country, and a class of the country, presented in the last column, denotes fertility rate. Altogether there are 77 basic attributes and 137 countries in our case study. There are 12 binary attributes and the rest numerical. Attributes and their values were partially obtained from the demographic sources such as UN [http://esa.un.org/unpp/] and Eurostat [http://epp.eurostat.ec.europa.eu/] databases and partially from Wikipedia.

In order to get relevant and appropriate explanations, data has to be manipulated in different ways and treated from different perspectives, with some strategic or tactical plan behind it (see Figure 1). Semantics behind the investigated phenomena should best be defined by the field expert. Data are usually manipulated on the basis of particular subgroups of learning examples, different class value arrangements or number of included attributes. In our case attributes were joined into 7 subgroups, based on different previous demographic theories: all (77), general country data (13), economical (11), social (10), educational (16), country health state (6), women's status in the country (39), and automatically selected by the Weka program (Witten & Frank, 2005). Our measurements were performed on all attributes and separately on specific groups like economical, consisting of 12 attributes such as unemployment rate, GDP ($) per habitant, GDP growth (%), etc.. For the basic class we have chosen Total Fertility Rate (TFR), discretized into two values: high (>2) and low (<2). The branching point 2 was chosen because it represents the replacement level of the population. In reality, replacement level is a bit higher, around 2.1, but this number depends on several other parameters such as mortality rate and immigrations, and furthermore only two countries have fertility rate between 2 and 2.1. Nevertheless, we also split the class into three values (high TFR: >3, middle TFR: 2<TFR<3, and low TFR: <2), but due to a lack of space only the results of the first version are presented here.

Further, we conducted our procedure separately on developed countries. Developed countries are countries with high gross domestic product (GDP); above 1000$ per habitant (38 countries). GDP is defined as the total market value of all final goods and services produced within a given country or region in a given period of time (usually a calendar year) (Sullivan & Sheffrin, 1996).

4.1.2 Data mining procedure

Our research group has decades of experience in developing and using data mining (DM) and machine learning (ML) systems such as Weka (Witten & Frank, 2005) and Orange (Demsar & Zupan, 2005), the latter being developed in our broader research group.

From the ML and DM techniques available in Weka and Orange we have chosen J48, the implementation of C4.5 (Witten & Frank, 2005), a method used for induction of classification trees. This method is most commonly used when the emphasis is on transparency of the constructed knowledge. In our case this was indeed so, since the task was to extract the most meaningful relation from hundreds of constructed trees.

The most meaningful relations are those most significant to humans with the best classification accuracy at the same time. To estimate the accuracy of the trees, we used a 10-fold cross-validation, built into the system. The estimated accuracy of a classification tree corresponds to a probability that a new example will be correctly classified.

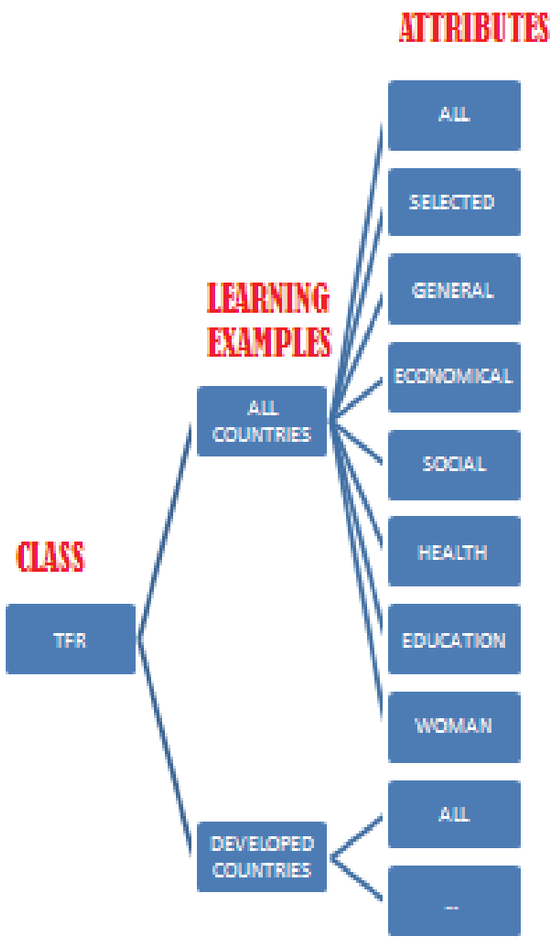

Fig. 1. The structure of the data manipulation.

We modified Weka´s J48 default parameters with the respect of the minimal number of objects in the nodes. We had long experienced the dilemma whether the constructed trees can be trusted. The standard DM approach was particularly challenged on the demographic problem since the relations are quite complex and need in-depth understanding. The decision which tree was the most useful one was based on the accuracy and quality estimations as described in the Section 3.1.2.

5. Findings and discussion

Tens of trees were created in a systematic way, as presented in Figure 1. First experiments were performed with all and then separately on the developed countries only. Finally, several selections of the attributes were tested: all, selected, general, economical, social, health, education and women. These tests resulted in 72 basic trees for all and 72 for developed countries only (8 different subgroup condition and 8 different possibilities of Minimum number of objects in the tree leafs). In addition, various further experiments with different class values were performed. In this section only the most interesting trees presented in Figure 1 were analyzed, i.e. those with most meaningful relations to humans and with best classification accuracy and quality at the same time.

		SUBGROUPS							
		All	**Selected**	**General**	**Economical**	**Social**	**Health**	**Education**	**Women**
ALL COUNTRIES	Attributes	L,Az, Ak,Ba	L,Az,Ak, 32,De	M	**Ae,Be,Di,V**	P,Ak,Ba	Az,Av	Ct,Cz,Cr	32,4,34, 37,19,6
	Min.Nr.Obj	3	**4**	**10**	**3**	3	8	7	**2**
	Acc (%)	81.0219	**83.2117**	**81.7518**	**83.9416**	81.0219	77.3723	80.292	**81.0219**
	F-measure	0.81	**0.834**	**0.817**	**0.841**	0.806	0.775	0.801	**0.811**
	AUC	0.817	**0.875**	**0.793**	**0.791**	0.809	0.836	0.803	**0.871**
	Kappa	**0.5939***	**0.6505***	**0.608***	**0.6644***	**0.5805***	**0.5216***	**0.5697***	**0.5971***
DEVELOPED	Attributes	37,L	20	An,Ap,N, Df, Af, K	FSI,u,Bh	**Ah**	//	Ce,Cs	37,25,26
	Min.Nr.Obj	9	2	4	2	**3**	//	5	4
	Acc (%)	81.5789	84.2105	68.4211	68.4211	**84.2105**	//	81.5789	68.4211
	F-measure	0.795	0.79	0.663	0.663	**0.831**	//	0.733	0.699
	AUC	0.541	0.438	0.461	0.417	**0.507**	//	0.542	0.624
	Kappa	0.2652	0.2138	-0.1813	-0.1813	**0.4093***		0	0.0539

* Quality of the conducted tree is acceptable

Table 1. Most relevant trees, induced from 8 different attributes subgroups, 9 different values of the parameter "Minimum number of objects in the tree leafs" processed on two different instance selections described with included attributes, minimum number of examples in the leaves and measures of accuracy and quality estimation.

Table 1 represents properties of the best trees constructed under given conditions. The attributes in the constructed tree are described here:

Ae =Human Development Index (HDI)
Af= The proportion of the adult population infected with HIV / AIDS (%) (2001-2003)
Ah = Prevalent Muslim religion
Ak = Legal abortion
An= Proportion of urban population (%)
Ap= The predominant race
Av= Prvate helath expaditure (% od BDP) (2003)
Az = The proportion of births attended by trained personnel an extensive (%) (1994-2004)
Ba= Proportion of married women (between 15 and 49 years) who use contraception
Be = GDP (Gross Domestic Product) Growth (%)
Bh= Exports of goods and services (% of GDP)
Ce= Public expenditure
Cr = Gross Enrolment Ratio. Pre-primary.
Cs= Gross Enrolment Ratio. Primary
Ct = Gross Enrolment Ratio. Secondary.
Cz = Pupil-teacher ratio. Primary
De = Internet users (per 1000 people)
Df= Personal computers (per 1000 people)
Di = GNI (Gross National Income) per capita
FSI= Failed States IndexK = Gosota (prebivalci na km^2)
L = Nr. Of stillborn children/ 1000 births
M = Literacy (%)
N= Life expectancy (male)
P = Punishable homosexuality
U = Unemployment (%) (2006)
V= GDP ($) per capita (2002-2007)
4= Percentage employees (women)
6 = Percentage employers (women)
19 = Youth (15-24) literacy rate (women)
20= Percentage of parliamentary seats in Single or Lower chamber occupied by women (2010)
25 = Women's share of tertiary enrolment (%)
26 = Female teachers, Primary education (%)
32 = Net enrolment ratio in secondary education (girls)
34 = Girls' share of secondary enrolment (%)
37 = Girls' share of primary enrolment (%)on education as percentage of GDP

In the case where all countries were taken into consideration, all groups of attributes provided trees of good quality, while when the experiment was processed only on developed countries, only the tree conducted from the social attributes seems to be significant. This means, that there are many factors or combination of them, that distinguish among countries with different TFR, while among developed countries, it is not clear, what is really connected with TFR of particular country.

5.1 All countries

What distinguish countries with lower TFR from the countries with higher TFR most, will be described with the following classification trees. The numbers in the leaf of the trees denote: class, the numbers of objects of the majority and minority class. Most representative classes are marked with a bolded frame.

The first obvious factor connected with TFR is economical situation and policy. From the tree on the Figure 2 we can see that countries with low HDI (lower than 0.771) usually has high TFR. On the other side, countries with high HDI mostly have low TFR, except if their GNI per capita is middle, GDP Growth (%) is low and BDP ($) per capita (2002-2007) is low. In the former case, the TFR is low as well. Over all we might conclude that developed countries have lower TFR in comparison with countries in development.

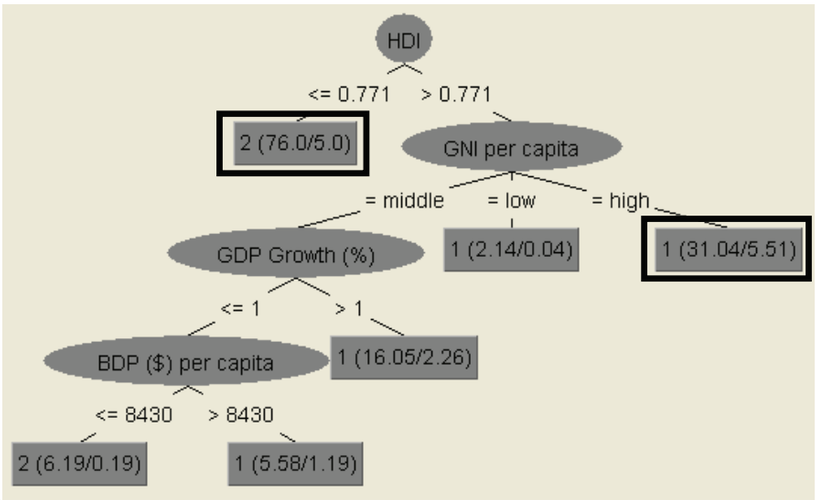

Fig. 2. Classification tree, induced from the economical attributes. Its classification accuracy is 83.9416 (F-measure: 0.841, ROC Area: 0.791, Kappa: 0.6644), the highest of the trees constructed on all countries.

Subgroup of general attributes also shows that countries development in terms of literacy is an indicator of TFR. As shown in the Figure 3, higher percentage of literacy is connected with lower TFR.

The tree constructed from social attributes also turned out to be of good quality (see Figure 4). It states, that TFR is low in the countries where abortion and contraception are allowed. The exceptions are the countries with illegal homosexuality. Whereas where abortion is not legal, high TFR is in the countries that also have high proportion of married women (between 15 and 49 years) who use contraception. Suggestions also appeared in the direction that Anti discrimination law is an indicator of low TFR countries. Overall impression is that more conservative countries have higher TFR.

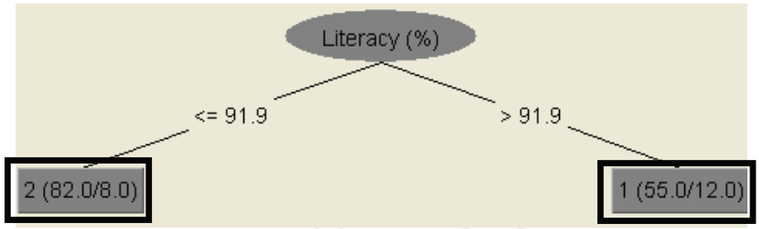

Fig. 3. Classification tree, induced from the general attributes. Its classification accuracy is 81.7518 (F-measure: 0.817, ROC Area: 0.793, Kappa: 0.608).

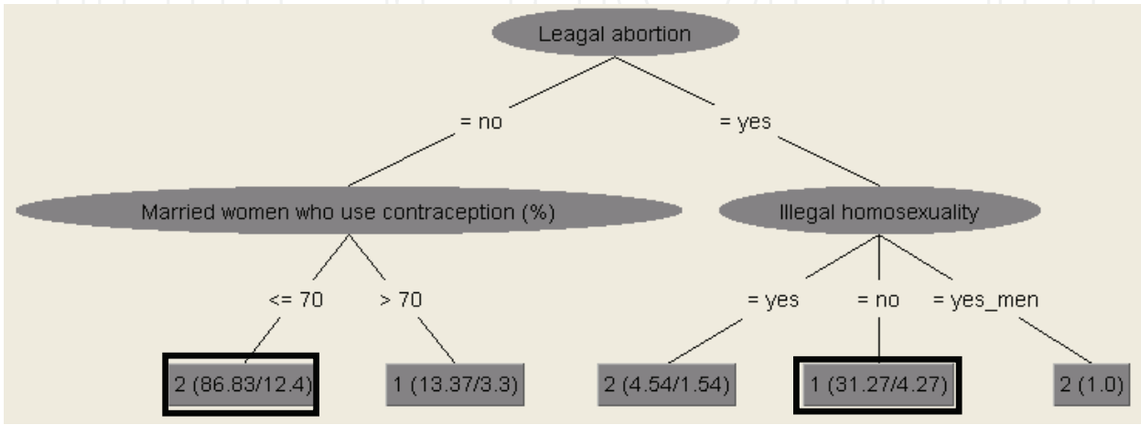

Fig. 4. Tree conducted from social attributes with the classification accuracy 81.0219 % (F-measure: 0.806, AUC: 0.809 and Kappa: 0.5805).

Among the most informative trees is the tree conducted from the attributes, automatically selected with Weka algorithm (see Figure 5).

What we see is that small number of stillborn children per 1000 births is an important indicator of lower TFR of the country. On the other hand, TFR is low when the proportion of births attended by trained personnel is extensive (more than 84%) and when abortion is legal. If abortion is not legal, than country will likely have low TFR when Net enrolment ratio of girls in secondary education is high (more than 74.5%) and internet users (per 1000 people) is more than 360. In general, one may say that where the countries health care is good, abortion is allowed, or at least where the country encourage women at their education and many citizens have an access to the internet, the TFR is lower than in the countries, which act in the opposite way. In comparison to the economic attributes, again the general impression is that the more one country is developed, the lower is TFR.

The status of the women in the country revealed a great importance of it. The most accurate tree that appeared from this iteration is the following one in Figure 6:

The tree shows us, that lower net enrolment ratio of girls in secondary education manly leads to higher TFR. Net enrollment ratio is the ratio of children of official school age based on the International Standard Classification of Education 1997 who are enrolled in school to the population of the corresponding official school age. The exceptions are the countries with high percentage of women employees. If on the other hand, net enrolment ratio of girls in secondary education is low, TFR is most of the times high. The biggest exception is the case, when Girls' share of secondary enrolment is high (more than 46.7%) and Girls' share of primary enrolment is low (less than 48.78%). The overall impression is that lower TFR is connected with higher involvement of women in education, especially in the higher education (on the secondary level or higher).

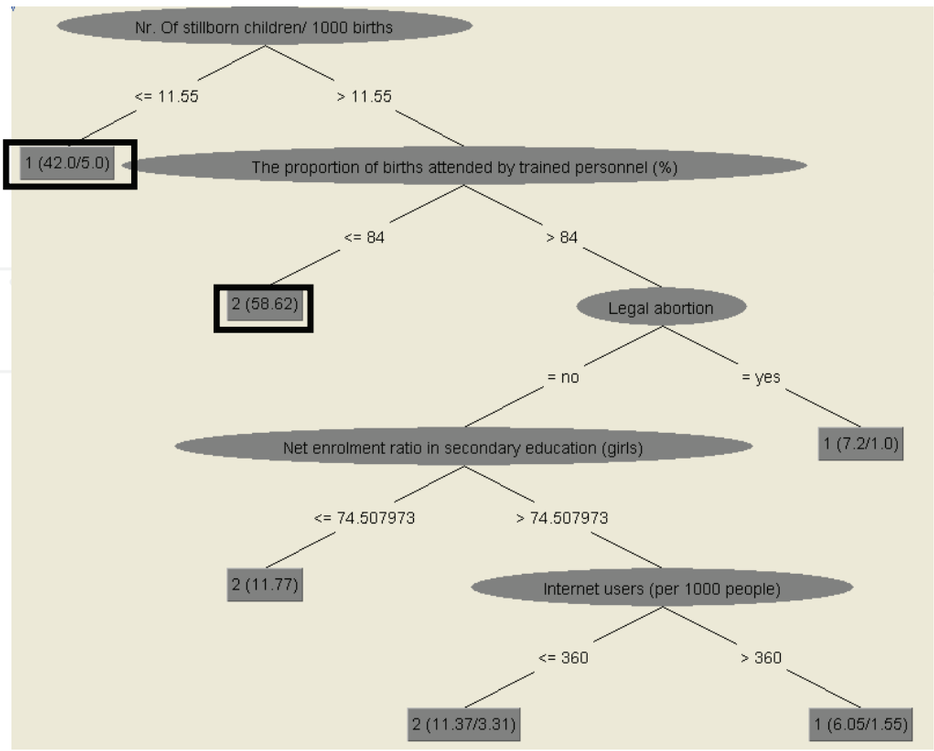

Fig. 5. Tree derived from Weka CfSSubsetEvaluation algorithm, with the accuracy 83.2117 (F-measure: 0.834 , AUC: 0.875 and Kappa: 0.6505).

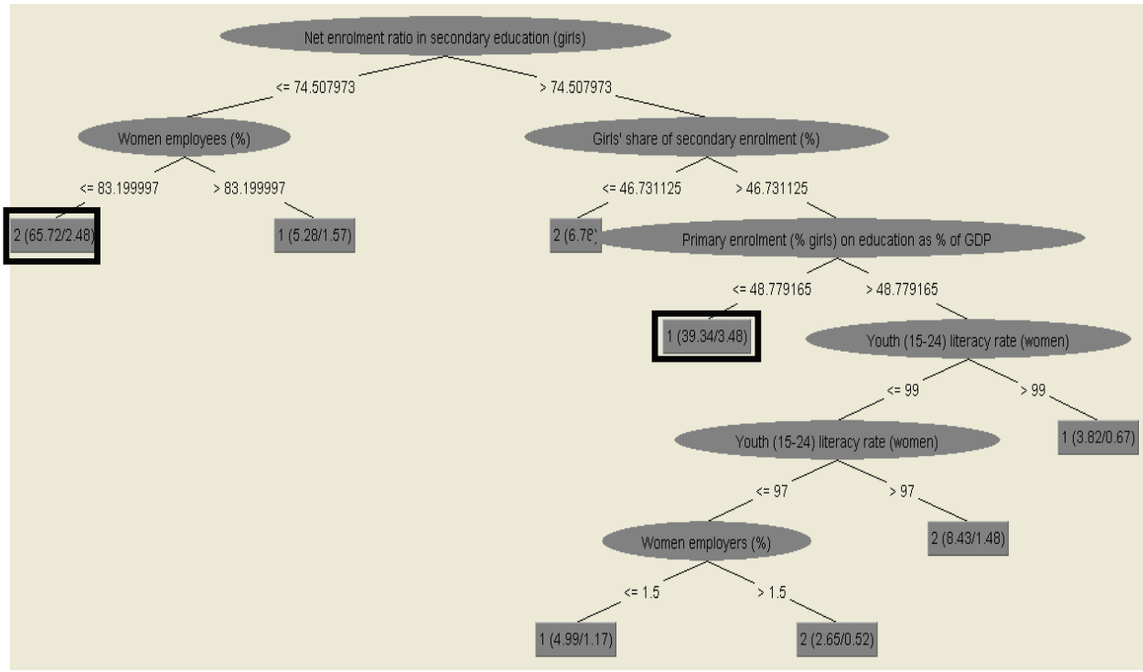

Fig. 6. Classification tree, induced from the attributes showing women status in the country. Its classification accuracy was 81.0219 (F-measure: 0.811, ROC Area: 0.871 and Kappa: 0.5971).

Are our latest findings indicator, that women career doesn't go hand in hand with family formation? Hakim proposes a theory, named Preference theory (Hakim, 2000). Preference theory seeks both to explain and predict women's choices regarding investment in productive or reproductive contributions to society. Preference theory is a historically-informed, empirically-based, multidisciplinary and predictive theory about women's choices between market work and family work. It proposes that there are three "qualitatively" different types of women, who differ among one another in their preferences about work and home: (a) the home-centred, who prefer a home life to labor market work, (b) the work-centred, of whom many are childless and all have strong commitment in their employment careers, and (c) the adaptive, who want to do some labor market work but do not commit themselves to their careers. She maintains that most women in modern affluent countries have genuine and unconstrained choices to choose between a home-career and a work-career according to their preferences, and therefore their preferences determine their home life and career. This has to do with changing gender roles. Now, young women wish to have other roles in life than that just be a mother. They seek a social status based on jobs they themselves hold and on the related financial rewards such jobs provide. Education has made them conscious of their capability; they want a just return from their years of schooling; and they wish to be considered as a autonomous individuals (Vitalli et. al., 2007) Although some authors claim that coinciding with the sharp reduction in fertility across the OECD (Organization for Economic Co-operation and Development), the correlation between fertility and female participation (and employment), which was negative during the 1960s and 1970s, became positive after 1986. From that year onward, fertility rates indeed slightly recovered in those countries with higher female participation rates whereas they suffered a sharp decline in those with low participation (Adsera, 2004).

5.2 Developed countries

As countries developmental status seems to be an important factor correlating TFR, we further took into consideration only developed countries. When doing this, economical factors lost their power of TFR prediction (tree accuracy is the lowest among all: 68.4211, F-measure, 0.663 and ROC Area: 0.417). The only statistically meaningful trees were those obtained from social attributes (See Figure 7).

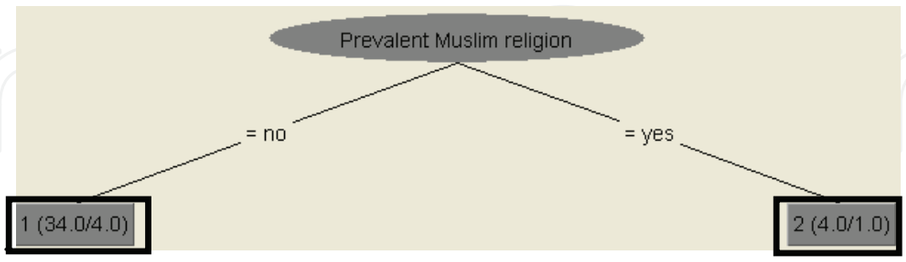

Fig. 7. Classification tree obtained from social attributes. (Accuracy: 84.2105%, F-measure: 0.831, ROC Area: 0.507 and Kappa: 0.4093).

The tree shows that prevalent Muslim religion is present in the developed countries with higher TFR. Further analysis should follow, investigating what the distinctive properties of developed Muslim countries are.

When removing the attribute showing prevalent Muslim religion in the country from the data the DM still produced a qualitative and interesting tree. Again it considers the religion;

if the prevalent religion is Christian, the TFR is low. If not, then the TFR is low if the religion is not official.
We can conclude that in developed countries religion is an important factor connected with TFR (Christianity with low TFR and Muslim with high).

5.3 General findings

We have found that the global population trends are rather uneven, as in the developed world population is now more or less stagnant and still under the regeneration limit, while in the less developed coutries population grows. The results confirmed the assumptions of the demographers; fertility is heavily influenced by the development stage of the country. As a developmental issue the literacy rate, contraception and health care situation are the pointers of fertility rate as well. The worse health care, no contraception and lower percentage of the literacy accompany high TFR. Furthermore, fertility rate is also connected with social policy of the county; namely, more conservative countries (illegal abortion, not allowed homosexuality, no antidiscrimination law) tend to have higher fertility rate than more liberal countries. Besides, an important aspect of fertility is situation of women in the country. It seems like the education of the women (specially the secondary and tertiary enrolment) and higher percentage of employment of the women is in connection with lower fertility. These results suggest that women's career doesn't goes hand in hand with establishing a family. Last but not least, religion turned out to be an important factor when considering fertility of the developed countries. The prevalent Muslim religion is connected with higher fertility, while Christianity with lower fertility rate.
When speaking broader than just about pure content of the trees, the following observations may be pointed out:

- Firstly, different measures of accuracy highly correlate among themselves. However, the measure of the tree significance is important. Namely, the relations in the conducted tree might easily be due to the coincidence – e.g. these with Kappa less than 0.5.
- After, subgroups that provided the most qualitative patterns discriminating countries with low from those of high fertility rate were examined further in order to establish that the appeared attributes are indeed the most valuable ones.
- Finally, it should be necessary to provide a detailed study of each particular attribute. Especially if the cause or a consequence relationship is to be discovered.

Regarding the fertility relations, ML tools enabled rediscovery of major well-known properties and they also provided several new ones. The case study of the demographic domain again proved that ML techniques are useful tool for mining social events.

6. Conclusion

The fertility analysis is just another field where data mining tools again proved their major asset: the constructed knowledge is in a transparent form, enabling human comprehension of relevant relations in complex forms and getting a holistic view of particular problem. In this way, an interactive and interaction process is enabled between computers and humans, exploiting the best properties of the two most advanced information machines. Regarding the fertility relations, the ML tools enabled rediscovery of major properties and provide several new comprehensions, sometimes even confronting general demographic knowledge. The case study demonstrated major advantages, threats and dilemmas:

- The new ML techniques enable exploiting several viewpoints thus enabling advanced understanding of the relations, their contexts and the actual weight instead of just getting transparent decision trees.
- When working with ML techniques we must be cautious on the quality of the results (rules, trees, etc.). Namely, not every tree or rule one get is significant enough. There is always a possibility of coincidence effect, which we check with the quality measures such as Kappa statistic.
- Is it indeed possible that there is a gap between the expert knowledge and the relations provided using ML techniques as showed with some contra intuitive relations provided by ML techniques?

At this point, it seems that the new methods might indeed be valid and that their "cold engineering" logic without social interplay can point out demographic relations in a new light, while we humans are so involved in subjective beliefs and wishes that might mislead us. Further analyses are needed to improve confidence in these tentative conclusions.

Once again, when using ML techniques for mining social events, these are the tree things one has to have in mind: meaning, accuracy and significance of the results.

New promising and extensive methods appear lately. One of them is Argument based machine learning (ABML) (Mozina et. al., 2007). It is a novel approach to machine learning, where classical machine learning is extended with concepts from the field of argumentation. This approach combines machine learning with explanations provided by domain experts. ABML is machine learning extended with certain concepts from argumentation. Arguments in ABML are a way to enable domain experts (in our case demographers) to provide their prior knowledge about a specific learning example that seems relevant for this case. In this case an experts' knowledge is not useful only for the explanation when the results are already obtained, but with the help of arguments and explanation it actually guides the procedure of building decision trees. The gained results are thus even more informative in the sense of explanation. It has already been successfully used in many domains such as chess, law and medicine (ABML) (Mozina et. al., 2006).

7. References

Adsera, A. (2004). Changing Fertility Rates in Developed Markets. The Impact of Labor Market Institutions. *Journal of Population Economics*,Vol.No.17:, January 2004, 1-27.

Angehrn, A. A. & Gibbert, M. (2008). *Learning Networks - Introduction, Background, Shift from bureaucracies to networks, Shift from training and development to learning, Shift from competitive to collaborative thinking, The three key challenges in learning networks.* The 1911 Encyclopedia Britannica.

Beitzel. S. M. (2006). *On Understanding and Classifying Web Queries. Phd Thesis*. http://ir.iit.edu/~steve/beitzel_phd_thesis.pdf.

Billari, F.C.; Furnkrantz, J. & Pskawetz A. (2000). Timing, sequencing, and quantum of life course events: a machine learning approach. Working paper 010, Max-Plank-Institute for Demographic Research, Rostock.

Christenson, M.; McDevitt, T.,; & Stanecki, K. (2004). *Global Population Profile: 2002. International Population Reports.* Health Studies Branch, International Programs Center, Washington Plaza II, Room 313A U.S. Census Bureau, Washington, DC 20233-8860.

Demsar, J. & Zupan B. (2005). From Experimental Machine Learning to Interactive Data Mining, White Paper (www.ailab.si/orange), Faculty of Computer and Information science, University of Ljubljana.

Fawcett , T (2006). An introduction to ROC analysis. *Pattern Recognit Lett,* Vol.No.27:861–874.

Gams, M. & Krivec J. (2007). Analiza vplivov na rodnost (Analysis of Impacts on Fertility). In J. Malačič, M. Gams (Eds.), *Proceedings of the 10th International Multi-conference Information Society (volume B) Slovenian Demographic Challenges of the 21st Century,* pp. 35-37. Ljubljana: Jožef Stefan Institute

Hakim, C. (2000). *Work-Lifestyle Choices in the 21st Century: Preference Theory*, Oxford Universit yPress

Hanley, J.A. & McNeil, B.J. (1983). "A method of comparing the areas under receiver operating characteristic curves derived from the same cases". *Radiology* 148 (3) 1983-09-01: 839–843. PMID 6878708.

Ivanov, K. (1972). *Quality-control of information: On the concept of accuracy of information in data banks and in management information systems*. The University of Stockholm and The Royal Institute of Technology. Doctoral dissertation.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 2 (12): 1137–1143.

Kandefer, M. & Shapiro, S. C. (2009). An f-measure for context-based information retrieval. In G. Lake- meyer, L. *Morgenstern, and M.-A. Williams, editors, Commonsense 2009: Proceedings of the Ninth International Symposium on Logical Formaliza- tions of Commonsense Reasoning*, pages 79–84, Toronto, CA. The Fields Institute, Toronto,CA

Landis, J.R.; & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33: 159–174.

Melville, P.; Yang, S.M.; Saar-Tsechansky, M. & Mooney R. Active learning for probability estimation using Jensen-Shannon divergence. *In Proceedings of the European Conference on Machine Learning (ECML),* pages 268–279. Springer, 2005.

Mitchell, T. (2006). *The discipline of machine learning* (Technical Report CMUML-06-108). Carnegie Mellon University.

Mozina,M.; Zabkar, J.; & Bratko,I. (2004). Implementation of and experiments with ABML and MLBA. ASPIC deliverable D3.4.

Mozina,M.; Zabkar, J.; & Bratko,I. (2007). Argument Based Machine Learning. *AI Journal*. Vol.171, No.10-15, July-October 2007, Pages 922-937

Obuchowski N.A. (2003). "Receiver operating characteristic curves and their use in radiology". Radiology 229 (1): 3–8. PMID 14519861.

Pazzani, M. (1991). The influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory & Cognition,* Vol. 17, 416–432

Sullivan, A. & Sheffrin S. M. (1996). *Economics: Principles in action.* Upper Saddle River, New Jersey 074589: Pearson Prentice Hall. pp. 57, 305. ISBN 0-13-063085-3.

Vidulin, V. & Gams M. (2006). Vpliv investicij v izobraževanje in R&R na gospodarsko rast, *Elektroteh. vestn.,* Vol. 73, No.. 5, pp. 285-290.

Vitali, A.; Billari, F.; Prskawetz, A. & Testa, M. (2007). Preference Theory and Low Fertility:A Comparative Perspektive, *European Demographic Research Papers*, No. 2, Vienna, Institute of Demography

Witten, I. H. & Frank E. (2005). Data Mining – Practical Machine Learning Tools and Techniques (sec. ed.), Morgan Kaufmann.

Knowledge-Oriented Applications in Data Mining
Edited by Prof. Kimito Funatsu

ISBN 978-953-307-154-1
Hard cover, 442 pages
Publisher InTech
Published online 21, January, 2011
Published in print edition January, 2011

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by 'Data Mining' address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Krivec Jana and Gams Matjaz (2011). Data Mining Techniques for Explaining Social Events, Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-154-1, InTech, Available from: http://www.intechopen.com/books/knowledge-oriented-applications-in-data-mining/data-mining-techniques-for-explaining-social-events

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen